Workshop Report on Shaping a Research and Policy Agenda on Big Data for Development in the Global South

Sriganesh Lokanathan and Thavisha Perera-Gomez

08 – 09 October 2016

LIRNE*asia*
info@lirneasia.net | www.lirneasia.net

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology*.
Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160. info@lirneasia.net
www.lirneasia.net

The Centre for Internet and Society (CIS) is a non-profit organisation that undertakes interdisciplinary research on internet and digital technologies from policy and academic perspectives. The areas of focus include digital accessibility for persons with disabilities, access to knowledge, intellectual property rights, openness (including open data, free and open source software, open standards, open access, open educational resources, and open video), internet governance, telecommunication reform, digital privacy, and cyber-security. 194, 2nd 'C' Cross, Domlur, 2nd Stage, Bengaluru, 560071, India. 080 4092 6283. sunil@cis-india.org

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada

IDRC | CRDI
International Development Research Centre
Centre de recherches pour le développement international

Canada

Acknowledgements

The workshop was jointly organizing by LIRNE*asia* and the Centre for Internet & Society and was funded by the International Development Research Centre, Canada.

This report was prepared by Sriganesh Lokanathan and Thavisha Perera-Gomez.

Table of Contents

Background and Objectives of the Workshop

LIRNE*asia* in partnership with the Centre for Internet and Society (CIS) convened a two-day workshop to discuss a 'research and policy agenda on big data for sustainable development in the Global South.' The workshop was held on 8th and 9th October, 2016 on the sidelines of the International Open Data Conference 2016 (IODC 2016).

The objective of the workshop was two-fold: (1) facilitate the development of a research and policy agenda on big data for development for the Global South, and (2) achieve a consensus on the way forward for a research network. The participants at the invitation-only workshop represented a cross-section of experts from policy/research institutes, civil society and industry that are active in big data work and research. A draft discussion document was circulated to the participants ahead of the workshop, covering some of the areas of discussion.

The workshop was opened by a welcome address by Sriganesh Lokanathan (LIRNE*asia*) and Elonnai Hickok (CIS) who set the set stage for the two days outlining the agenda and the purpose of the workshop. Subsequently, use case examples of big data for development were shared by some of the participants. The breakout discussion sessions focused on (a) Representativity & Marginalization (Gender, socio- economics, race, etc.) (b) Researching harms (Competition, Privacy, Security, Surveillance) (c) Researching solutions (Legislation, regulation, ethics) and (d) Challenges in big data for development (Research capacity, data, policy impact). CIS also shared the preliminary results from their exercise of mapping actors involved in big data for development in the Global South. The workshop concluded with a discussion on the way forward.

This document summarizes the presentations and discussions of each session and presents the key points from each discussion. Further information, such as the agenda, the presentation slides and contributing documents are available online at http://lirneasia.net/2016/10/workshop-on-shaping-a-research-and-policy-agenda-on-big-data-for-sustainable-development-in-the-global-south/.

Summary of Key Discussions

Applications of Big Data for Development

The objective of this session was to showcase some of the results of research using big data analysis for development purposes. Prof. Ryosuke Shibasaki from the University of Tokyo, Prof. Joshua Blumenstock from the University of California, Berkeley, Ms. Diastika Rahwidiati and Dr. John Quinn from the UN Global Pulse, and Sriganesh Lokanathan from LIRNE*asia* shared their experience in conducting research using big data analysis.

Mobile phone data and satellite data to predict poverty and wealth. Prof. Joshua Blumenstock shared the results of his research on leveraging mobile phone data and supervised learning to predict poverty and wealth, analyzing call detail records which were supplemented by phone surveys with the final results validated through household survey data. He also shared results of another study by Jean et al. (2016) on predicting poverty using satellite imagery and

machine learning whereby neural networks learned satellite imagery features that correlated with economic activity.

Human mobility patterns and dynamic census. Prof. Ryosuke Shibasaki showcased the use of GPS technology on mobile phones to understand mobility patterns after an earthquake, visualizing mobility in the period before and after the earthquake. Moreover, he spoke on the development of a dynamic census, estimating population demographics and trajectories based on CDR data, highlighting that this method would enable the capture of those in the base of the pyramid as well as population who may be difficult to reach through field surveys.

Sentiment analysis using big data sources. Dr. John Quinn and Ms. Diastika Rahwidiati cited several examples of UN Global Pulse's work in the space including converting public radio broadcasts into machine-readable form by using speech translation and speech recognition technology as well as a project to leverage satellite imagery and image processing software to identify and count thatched roofs (as a proxy-indicator of poverty). Two other prototypes that were highlighted were the Social Listener tool and the Haze Grazer tool with the former leveraging citizen feedback (passive) on social media as well as existing complaint systems to generate insights on citizen complaint.

Mobile data for urban planning and disease propagation. Mr. Sriganesh Lokanathan highlighted some of the ongoing big data analyses undertaken by LIRNEasia including understanding changes in population density in the Colombo region using CDRs, understanding the impact of the new expressways in Sri Lanka on travel patterns, understanding traffic conditions by coupling with CCTV footage, as well as working on building models that could understand the spatial propagation patterns of communicable diseases like dengue. Mr. Lokanathan also spoke of extensive engagement with policy domains and their symbolic environments to enlighten them.

Implications of big data to inform public policy. Participants heard of the potential of big data to provide interim statistics interval between official surveys. Similarly, mobile data can be used to generate population maps and mobility/migration patterns which are generally part of census and household surveys. Moreover, it would enable the near real-time monitoring of shocks and vulnerability in the economy, helping better evaluate policy impact.

Limitations of big data highlighted. The presenters also outlined limitations of big data use such as non-representativeness of mobile adoption and the potential for representation bias, data access and privacy, limitations of anonymized data among others.

Breakout Sessions

With primacy given for discussion, the workshop was structured around various small group discussions. During each small group session, a resource person spent around 15 minutes framing the discussion. Workshop participants were then divided into small groups with each group having an assigned rapporteur from among the participants. A time was set-aside in the agenda for each rapporteur to report back to the whole group. Discussions addressed both research using big data analysis as well as research on big data issues. The key points from each session have been outline below:

1.1.1. *Representativity and Marginalization*

The objective of this session was to identify key issues in representatively and marginalization in the analysis of big data with the discussion being framed around inter alia gender, socio-economics, race.

Key challenges in addressing gender in big data research. In addition to the issue of identifying gender from an anonymized or pseudonymized dataset, there were also questions of how address gender disparities in data access. Some potential solutions offered by participants included statistical corrections and full-immersion ethnography, although the fundamental issue of legibility would still remain, i.e. if people are not observed in the data, their voices won't be amplified because there is no voice to be found.

Better calibration of big data to address representativeness and marginalization. This included the establishment of guidelines and calibration (through surveys and/or other data). Participants also explored how such techniques could be validated. For example, it would important to ask a series of questions of particular big data analyses, to understand its representativeness. Similarly it would be important to develop methods of improving the interpretation of results with respect to its representativeness, testing the accuracy of predictive models and algorithms, as well as evaluate the type of frameworks that could used to explain use of big data.

Practical issues surrounding potential solutions for representativity and marginalization. The practicality of collecting ground truth data when there is a disproportionate male presence (which was often the case according to some experts) was considered, raising the issues around the cost-benefits of trying to assess representativity and whether it makes sense in every context. For example, would people have the patience to make the necessary corrections, conducting a thorough exercise of calibrating their big data estimates and survey-based data? Moreover, in addition to the time consuming nature of the task, the process itself would prove to be expensive – a constraint for those working with data.

The risks of relying on single use cases when addressing representativity issues. A strong focus on measuring the impact of a particular method on representativity in one particular context would not necessarily mean that method would be possible and/or relevant in other contexts, raising concerns about the distortionary effects of prioritizing a single use case.

1.1.1.1 *Researching Harms*

The objective of this session was to discuss the key focal points when researching harms of big data analysis, with the discussion being framed around issues of competition, security, surveillance and privacy.

Shifting from theorizing about harms to assessing impact of actual harms. It was noted that instead of theorizing about the various possibilities of harm arising from the analysis of big data, it was important to undertake case studies around issues of actual harm that have already happened. These would be useful to develop measures to limit the impact, and possibly prevent such harms, rather than development of such measures in the abstract. Conversely, there was an argument that it would

not be effective to wait for big data harms to actually take place, and for case studies to be then developed, before precautionary measures are taken towards preventing and limiting such incidents.

Ethical considerations when conducting big data analysis. Participants raised the ethical issue of big data being used outside of the primary purpose it was collected for. For example, a bank making decisions on providing credit/loans based on potential customers' mobile phone usage and/or airtime credit purchases.

The risk of big data solutions leading to unfair competition. The emergence of new platforms have given rise to entities that have data access because they control the platform. There is opportunity for such entities to use the insights generated from such data into other lines of business where the competitors for that business don't have access to that same information. Similarly, with regards to innovation, the ability of start-ups to function/not function because of potential lack of access to data was also raised.

Identifying risks of not using big data. While there are risks associated with using big data, there are also risks associated with not using the data as well. Thus, it would be important to conduct research that attempts to gauge the cost of not using [big] data in the developing world.

Risk assessment frameworks may be useful for addressing harms, but contextual variations could limit the effectiveness of a generic framework. Utilizing risk assessment methodology to manage data and also in applying data models/algorithms was explored. Participants highlighted the need to factor in both the likelihood of a harm happening as well as the impact of the potential harm along a continuum. Participants spoke in terms of size of harm and threat modeling and predicted severity of misuse, noting that these were all probabilistic assessment of harm. On the other hand, it was pointed out that risk assessment frameworks should be country specific given that what is an unacceptable risk in one country in not necessarily an unacceptable risk in another country. One way to address this was through conducting national surveys to inform policy makers and stakeholders of relevant risks – a time consuming and expensive task. A proposed alternative was to look at case studies and not do expensive surveys, so scarce resources could be spent on conducting research using big data analysis rather than on researching harms.

The inclusion and development of professional standards into big data analysis. The fact that professional standards are able to transcend jurisdiction enables at least de facto, to standardize and open discussion on responsible practices and what different actors need to consider at various stages. This leads to the role of codes of practice or regional codes of practice given that although the principles will be the same, there will be differences in regulation.

Conventional privacy solutions not applicable to many big data 'harms' and therefore ex-ante solutions are a better alternative. The conventional privacy solution, which is inform and consent, doesn't apply to a number of harms related to big data including for example, insecurity, disclosure, exposure and increased accessibility. Solutions could possibly include ex-ante solutions like standards, insurance, and Service Level Agreements (SLAs) with the service providers/data collectors that would lead to stronger security protocols/safeguards.

Investing in capacity building to address privacy harms. The need to engage in capacity building in relation to privacy was underscored in the discussions. In addition to building research capacity it was noted that data management capacities in both governmental and non-governmental agencies need

to be developed so that big data meant for development research is managed, stored, and used in a secure manner, and the raw data is exposed to an absolutely minimum number of users.

Awareness of local context a key feature in addressing privacy issues. Understanding the regulatory environment in which researchers operate was another point that was stressed, given that laws relating to the use of data vary by country. The benefit of a local partner was underscored as a way in which the contextual issue of privacy could be more practically addressed. The presence of a local partner in every team conducting research involving the analysis of big data would help ensure that local concerns and local context on what's acceptable and what's not could be reflected in the research, as an alternative to global frameworks being imposed on from the top.

1.1.1.2 Challenges and Researching Solutions

The objective of these two sessions was to identify key challenges in big data research and brainstorm potential solutions. The discussions were framed around legislation, regulation, ethics, research capacity, data access, and policy impact.

The danger of dilettante data science bridging the gap between big data research and policy. Participants explored the dichotomy between big data in research and policy arguing that although peer reviews tended to be flawed and methods were generally opaque, there was some form of oversight on research while the squeaky wheel drove policy action. The danger of dilettante data science filling this gap was highlighted, raising the argument as to whether the gap needs to be bridged or if it can be considered as merely growing pains.

Research capacity and lack of awareness of big data solutions key challenges in global south. The knowledge and skills necessary to conduct big data analysis for development purposes were multidisciplinary including but not limited to computer science, statistics and domain knowledge. In addition to building capacity on the research side, there is also the need to build capacity to absorb the results of the research. Policy makers and stakeholders in the symbolic environment of these domains need to be informed consumers of big data research so that they are able to ask the right questions regarding the veracity of the research results generated from big data analysis.

Well-articulated evidence-based solutions to inform policy makers of benefits of leveraging big data. Mapping the incentives of leveraging big data analysis along with examples of success stories for similar problems were other factors that were raised. In order to do this, it is important to understand the problem well and articulate it in the form of a story, helping policymakers rethink the way they see problems and positioning the researcher/research organizations as having the resources to be part of the solution. Participants also stressed the importance of providing policy makers tangible benefits along with potential results. Moreover, it would be beneficial to build on existing tools and customizing them to the local context to show the potential of leveraging big data.

Engaging users of results in the research process. Involving policy makers/stakeholders who will be using the results of the big data analysis in the research process would ensure that they are involved in the exploration, better positioning them to understand the insights derived. Tighter feedback loops would also contribute towards this. This would also enable the researchers to better understand the

policy relevance of their research and how their work would potentially be utilized, which helps researchers to ask better research questions.

Capitalize on opportunity for social good to attract new talent into the space. Competition for professionals in data science is fierce and researchers may not be able to pay competitive rates, however, what they can provide is the opportunity for social good and this ability to make an impact is something researchers can capitalize on when attracting new talent.

Establishment of a data experimentation incubator. Another option was the introduction of a data experimentation incubator that would act as a secure platform for researchers to collaborate. This could encourage startups to work together and for researchers to share best practices. The concept of a hub was also touched upon by participants whereby researchers could network, share learning and potentially collaborate on research engagements. This would lead to increased use cases.

Linking up with higher education institutions. Given the multidisciplinary nature of the work involved, it would also be helpful to link up with universities to develop a multidisciplinary research education environment in order to build capacity for analysis of data for development as well as to mentor students with the aim of hiring them when they graduate. Additionally, participants also proposed leveraging online learning platforms to advance knowledge to build capacities than merely relying on universities. Moreover, access to more open data would help students explore big data analysis.

Mapping Actors Engaged in Big Data for Development in the Global South

The Center for Internet & Society (CIS) and LIRNE*asia* sought to identify actors who are involved in and/or could be involved in the big data for development discourse globally, with a particular focus on the 'Global South. CIS presented some of the observations based on the preliminary results of their exercise.

Observations of mapping exercise. Some of the observations based on the preliminary results include (1) big data and rights dialogue can be brought into the big data and development work; (2) there is a need for grassroots initiatives around big data for development in the Global South that can draw insights from Global North institutions but are not driven by Global North institutions and (3) many actors in big data for development in the Global South are emerging, but there are only limited actors doing big data analytics for development. This presents an important opportunity for 'on the ground' networks to be formed.

The Way Forward

The objective of the session was to identify key ideas that were at the frontiers of research in big data for development and provide potential approaches for the development of a research agenda.

Features of emerging research frameworks. Participants explored whether emerging research frameworks should be practice-oriented or normative and whether there should be a differentiation for state and private actors. The question was also raised whether big data actually need a completely new framework with new mechanisms or whether existing frameworks could be adapted. Moreover, participants also discussed whether these new/adapted mechanisms should be developmental, protective or enabling and how an appropriate balance could be struck. Irrespective, the question

remain with regards to the capacity of governments (and especially those in the Global South) to create/ adapt such frameworks.

Capacity building to play a key role in proposed network. If the objective was to develop voices from the Global South, then capacity building was key. Thus Southern actors need to be able to conduct analytics as well as have deep contextual understanding of amongst others, marginalization, privacy, and competition issues.

Focus of network dependent on time horizon and investment by funder. The funder's commitment in terms of time horizon and investment is a key determinant of the research vs. policy focus of the network. Capacity building for research in big data and data science is a medium to long term investment with a pipeline that would allow for students in their undergraduate years to work for a couple of years in the analytics space (with a developmental focus). From there they could then work in industry or policy or continue their education. Thus if the time horizon was around three years, instead of building a research pipeline, it would make more sense to focus on policy. Moreover, the research angle could look more at research about big data for development and less on conducting big data analysis for development purposes.

Network to build connections. Another suggestion was the network to act as more of a hub connecting researchers and research organizations with one another, with the goal being collaborative research amongst atleast a sub-set of the members. Moreover, if there is research being conducted on the same issue in parallel, this initiative can ensure that there would be a process of osmosis between those two projects.

The conversations held over the two days will all serve as inputs to the funder on how a network could be formulated.

Proceedings

Session 1: What is big data for development and what are its uses?

Prof. Joshua Blumenstock spoke on data-intensive development as a new field that combines big data with machine learning and development theory. He outlined that many 'big' datasets common in many developed regions were still rare in other regions with two key exceptions: mobile phone usage data and satellite imagery. He covered the example of leveraging mobile phone data to predict poverty and wealth by supplementing call detail records and individual surveys and validating it with national household survey data. He also spoke on the usage of satellite imagery to estimate regional poverty by leveraging neural networks learned features in satellite images that correlated with economic activity. Thus, images from daytime satellite imagery were used to associate features from these images with nightlight luminosity enabling a stronger proxy for economic activity.

Prof. Ryosuke Shibasaki showcased the use of GPS technology on mobile phones to understand mobility patterns after an earthquake, visualizing mobility in the period before and after the earthquake. Moreover, he spoke on the development of a dynamic census, estimating population

demographics and trajectories based on CDR data, highlighting that this method would enable the capture of those in the base of the pyramid as well as population who may be difficult to reach through field surveys. He also addressed how some of the challenges of CDR data could be addressed through the dynamic census method. For example, irregular record interval by interpolation and anonymized CDR data by demographic attribute estimation, supplemented by mobile phone user survey data.

Dr. John Quinn and Ms. Diastika Rahwidiati cited several examples of UN Global Pulse's work in the space including converting public radio broadcasts into machine-readable form by using speech translation and speech recognition technology as well as a project to leverage satellite imagery and image processing software to identify and count thatched roofs (a proxy-indicator of poverty). Two other prototypes that were highlighted were the Social Listener tool and the Haze Grazer tool with the former leveraging citizen feedback (passive) on social media as well as existing complaint systems to generate insights on citizen complaint, and the latter which uses open data such as social media, citizen journalism videos, satellite imagery, Indonesia's national complaint system among others to offer situational information regarding fire and haze and public perception.

Mr. Sriganesh Lokanathan highlighted some of the ongoing big data research undertaken by LIRNEasia including understanding changes in population density in the Colombo region using CDRs, understanding the impact of the new expressways in Sri Lanka on travel patterns, understanding traffic conditions by coupling with CCTV footage, as well as working on building models that could understand the spatial propagation patterns of communicable diseases like dengue. Mr. Lokanathan further highlighted that key characteristics of these new data offers significant benefits for development purposes. Mr. Lokanathan also spoke on the implications of big data to inform public policy. He also spoke of extensive engagement with policy makers and symbolic environment to enlighten them.[1]

Session 2: Representativity & Marginalization

Prof. Bitange Ndemo framed the discussion on the implications of representativity and marginalization for big data for development research. He also spoke about the World Population Project (www.population.io) whose aim is to make demography data available to a broader audience enabling a greater understanding of economic and social development.

The breakout sessions saw rapporteurs collectively capture a variety of issues:

[1] More information about LIRNEasia's ongoing big data for development research is available from http://lirneasia.net/projects/bd4d/

The dichotomy between big data in research and policy was explored, with concerns being raised that although peer reviews tended to be flawed and methods were generally opaque, there was some form of oversight on research compared to policy action, which was driven by the squeaky wheel. This created the danger of dilettante data science filling this gap. Conversely, discussion revolved around the size of the gap between research and policy and whether it should be considered as growing pains. For example when considered in terms of the need for data-driven decisions, the gap was too big, but the gap could be considered small since policymakers don't appear to particularly care about assumptions/caveats of research.

When considering gender in the context of representativity here are questions related to understanding whether the data sets being analyzed were representative of gender. At the same time there are questions in relation to gender disparities in technology access that also need to be considered in the research. Some potential solutions offered included statistical corrections and full-immersion ethnography, although the basic issue of legibility would still exist, i.e. if people are not observed in the data, their voices won't be amplified because there is no voice to be found. Various forms of marginalization including age, gender, region, religion, access to broadband, and mobile phone/smart phones were identified. The issue of marginalization was also discussed in terms of the use and the user, i.e. who has access to the data, who is represented and how they are being represented as well as in terms of who is analyzing the data and how is it being analyzed. This raised questions on not just the characteristics of the data itself, but on the analysis of data and how to ensure that the algorithms that are being leveraged are not hidden.

There was further discussion how to better calibrate big data to account for representativeness and marginalization for example, through the establishment of guidelines, calibration methods and calibration data (which would most likely involve a comprehensive survey) and use cases and statistical data. Participants also explored how such techniques could be validated. For example, if there is a black box and the results derived from it will be utilized, then thinking through what questions to ask the black box to ensure representation, how can the interpretation of results be improved, how can the accuracy of predictive models and algorithms be tested and which kind of frameworks can be use to explain use of big data. The question was also raised whether data scientists can be forced to report assumptions, errors, confidence intervals as alternatives to enforce accountability and rigor

The need for to standards to be context/country-specific was underscored. Moreover, concerns were raised over the distortionary effects of prioritizing a single use case. For example, if there were a strong focus on measuring the impact of a particular method in one particular context, would that method be possible and/or relevant in other contexts? Concerns were also raised regarding the practicality of correcting for the disproportionate male presence and whether it makes sense in every context in terms of cost-benefit. For example, would people have the patience to make the necessary corrections, conducting a thorough exercise of rounding their big data estimates and survey-based data? Moreover, in addition to the time consuming nature of the task, the process itself would prove to be expensive – a constraint for those working with data.

Session 3: Researching harms

Prof. Rohan Samarajiva framed this discussion mainly leveraging work by Solove (2008) on privacy harms. This included information collection (surveillance), information processing (aggregation, identification, insecurity, secondary use and exclusion), information dissemination (disclosure, exposure, increased accessibility). He noted that the conventional solution to issues of privacy, which is usually inform and consent, did not apply to a number of the outlined harms. Potential solutions could include ex-ante solutions like standards, insurance, and service level agreements with service providers/data collectors. He also noted that while there was interest in group harms, rights are usually associated with individuals and not groups (although harms may occur to groups) with the only collective right recognized by international law being that of peoples' right to self-determination. He concluded by highlighting that the effects on competition were one of the most neglected aspects of big data.

Rapporteurs raised numerous issues in relation to the harms associated with big data. Concerns were raised around the use of the word 'harm,' specifically if there was a difference between harms associated with data in general and those associated with big data. Questions were also raised as to if a differentiation was required in differentiation harms according the level of risk and/or the level of unexpected/unintentional consequences. It was noted that instead of theorizing about the various possibilities for harm in utilizing big data, there was an urgent need to undertake case studies around actual occurrences of harm, which could then be used to develop measures to limit the impact of such harms in future and possible prevent it. Conversely, there was an argument that it was not realistic to wait for big data harms to actually unfold, and case studies to be developed, before precautionary measures are taken towards preventing and limiting such incidents. Moreover, there was also the argument that precautionary measures should not inhibit the reaping of benefits from the patterns and insights that can be obtained by analyzing big data.

The perspective from which harms were assessed was also important according to participants. For example, there are differences when considering the perspective of a party who has access to data and is looking to apply it in a particular circumstance, versus the perspective of framing research in third party/external actors. Big data harms were also discussed in terms of extent of its impact, and the predicted severity of misuse and/or abuse with the addition that these were all a probabilistic assessment of harm. The issue of where in the lifecycle of harm, the concepts being discussed fit in to was also discussed. It was also noted that the value to humanitarian/development efforts would play a crucial role in gaining access to data, however difficulties in defining/evaluating such value would pose a challenge. Moreover, while the cost-benefit analysis tended to be use-based, the benefits were often indeterminate. There was also discussion on the effects of big data on competition, particularly small businesses, which had limited capacity to absorb and use big data. This was highlighted as equity concerns that any

government has to deal with, but at the same time these may also be market opportunities for new businesses that could fill the analytical gap for small businesses.

One suggested way to address big data harms was through risk-assessment frameworks. The discussion focused on how organizations could apply risk assessment methodology to in aspects related to basic data management, as well as when applying data models/ algorithms. These could be applied from a process perspective e.g. focused on security and risk/ liability (threat modeling) during collection, transmission, processing, storage, publication, and retention. It could also be substantive including a review of embedded assumptions, flaws or mitigating (unaccounted for) factors, etc. It was noted that each of these assessments should factor in the likelihood of a harm happening as well as the impact of the harm along a continuum. Also discussed was the impact of dependencies between actors or the intentions of the original terms of acquisition on data evaluation standards, highlighting that institutions that are unaffiliated to researchers, users and/or consumers would be better positioned to conduct harm/risk assessment. References were also made to the UN Global Pulse risk assessment framework.

At the same time there was a countervailing argument that risk assessments frameworks were not effective given the general perception that 'risk is bad' when compared to the neutral sense of the word. Moreover participants raised the point that risk assessment frameworks should be country specific given that what is maybe considered unacceptable in one country might not be in another country. One way to address this was through conducting national surveys to inform policy makers and stakeholders of relevant risks – a time consuming and expensive task. A proposed alternative was to look at case studies rather than expensive surveys, so that scarce resources could be spent on conducting research rather than researching harms.

The need to engage in capacity building in relation to privacy was underscored in the discussions. In addition to building research capacity it was noted that data management capacities in both governmental and non-governmental agencies need to be developed so that big data meant for development research is managed, stored, and used in a secure manner, and the raw data is exposed to an absolutely minimum number of users. Understanding the regulatory environment in which they operate was another point that was stressed given that that some countries have very tight laws in relation to using data whilst others have very light to non-existant regulations. The inclusion of professional standards in this brand of analysis was also brought into the picture given that professional standards are able to transcend jurisdiction enabling at least de facto, to standardize and open discussion on responsible practices and what different actors need to consider at various stages. The benefit of a local partner was underscored as a way in which the contextual issue of privacy could be more practically addressed. The presence of a local partner in every big data research team would help ensure that local concerns, local context could be brought into the work without global frameworks being imposed on from the top.

Session 4: Researching solutions

Elonnai Hickok framed this discussion around legislation, regulation and ethics. This required understanding the gap between what is possible, but not legal; what is legal and possible; as well as what is possible but not regulated. She raised questions on the adequacy of regulation and legislation as well as the adequacy of ethical frameworks and ethical practices. She posed some questions/issues to the audience including: (1) what are the criteria to be considered when assessing the ethics of collection, analysis, and use in the context of big data, (2) technical application of big data techniques (3) are there specific ethical considerations in using big data in emerging economies (4) do the ethics question and regulation question change in the context of the state vs. the private sector (5) privacy regulation for the public sector and private sector.

Participants explored whether emerging research frameworks should be practice-oriented or normative and whether there should be a differentiation for state and private actors. Moreover another point brought up for consideration was whether there was a need for a new set of mechanisms versus adapting existing frameworks and if the latter case, what was the capacity of governments to create these new mechanisms?

Moreover, participants also discussed whether these new or adapted mechanisms should be developmental, protective or enabling and how an appropriate balance could be struck. The need for a cost-benefit analysis framework was also touched upon. In terms of frameworks, another area that was highlighted was the overarching nature of ethics and what needs to be considered when looking at it. There is also the question of whether data scientists, who are the ones working on the data should be in control of the data, and decide what ethical framework to use.

There was also a discussion around whether access to data held by private entities for public sector be based on legislation/regulation or through contractual mechanisms and if the latter, the types of contracts that were needed and their respective price points. Participants suggested possible external accreditation for the work that is done in the big data field, for example through professional standards bodies. This leads to the role of codes of practice or regional codes of practice given that although the principles will be the same, there will be differences in regulation.

The advent of new data visualization tools brings new opportunities for presenting data in a manner that the general public can understand. Moreover, sharing information with the public may also be a matter of ethics for example, “is it ethical to conduct an analysis conducted to understand what is happening with the poor people and there isn’t a simple visualization to present the findings with those whose information were used in the analysis?”

Moreover, while there are risks when associated with using big data, there are risks associated with not using the data as well. Thus, it would be important to try to find out the cost of not using [big] data in the developing world. Creating and retaining capacity is a challenge in the global south. Those with knowledge and skills are attracted by the private sector. Participants discussed ways to attract relevant skilled personnel by focusing on the impact of their work. Moreover, the

contribution of open data to developing skills was explored as well as the establishment of hubs to share expertise and connect with others in the field.

Participants also touched on the power dynamics existing between the global north and the global south as well as the dynamics that exist between people who are analyzing the data and the data user, raising the questions about rights. Thus, power and rights would be thrown into a discussion around ethics, regulation and legislation as well.

DAY TWO

Session 1: Challenges in big data for development

Sriganesh Lokanathan framed this discussion and spoke about some of the key challenges in big data for development and the need to address them if we are to capitalize on big data for development research in the Global South. The lack of skills and the need to build research capacity were highlighted as key considerations for the Global South. The knowledge and skills necessary for big data for development roles were multidisciplinary including, but not limited to computer science, statistics and domain knowledge. In addition to building capacity on research, there is also the need to build capacity to absorb the results of the research. Another key challenge he outlined was gaining access to data. This encompassed reducing transaction costs of data access, implications on competition by gaining access to privately held data.

Through the group discussions, participants proposed a range of solutions to address identified challenges.

In terms of creating demand for big data for development solutions, participants proposed mapping the incentives of leveraging big data analysis in a manner that is easy to comprehend, along with examples of success stories for similar problems.. In order to do so, it is important to understand the problem well and articulate it in the form of a story, helping policymakers rethink the way they see problems and positioning the research entity as having the resources to be part of the solution. In other words, strong articulation of the problem with evidence of how it can be solved. Moreover, participants proposed other ways of data dissemination including short 'Ted Talk' style videos that were designed for policy makers that demonstrates the value of big data for development. Other suggestions included seminars and workshops.

Participants also stressed the importance of providing policy makers tangible benefits along with potential results. Moreover, it would be beneficial to build on existing tools and customizing them to the local context to show the potential of leveraging big data. Furthermore, involving policy makers/stakeholders who will be using the results of the analysis in the research process would ensure that they are involved in the exploration stage, better positioning them to understand the

insights derived. Tighter feedback loops would also contribute towards this. Moreover, understanding the purpose of the research, that is, understanding the uses of the final result would help researchers ask better questions and better articulate the proposed solution.

Given that attracting and retaining talent were key constraints in the big data for development space, participants explored various solutions to address this. They identified hackathons and other similar events as a means to help spot talent. Competition for professionals in data science is fierce and researchers may not be able to pay competitive rates, however, what they can provide is the opportunity for social good and this ability to make an impact is something researchers can capitalize on when attracting talent.

Another option was the introduction of a data experimentation incubator that would act as a secure platform for collaboration for researchers. This could encourage startups to work together and for researchers to share best practices. The concept of a hub was also touched upon by participants whereby researchers could network, share learning and potentially collaborate on research engagements and lead to increased use cases.

Given the multidisciplinary nature of the work involved, it would also be helpful to link up with universities to develop a multidisciplinary research education environment in order to build capacity for analysis of data for development as well as to mentor students with the aim of hiring them when they graduate. Additionally participants also proposed leveraging online learning platforms to advance knowledge to build capacities than merely relying on universities. Moreover, access to more open data would help students explore big data analysis.

Session 2: Mapping actors for big data for development

Elonnai Hickok and Sumandro Chattapadhyay from CIS presented some of their observations based on the preliminary results of their exercise on systematic mapping of big data for development stakeholders with a focus on the Global South in which CIS and LIRNEasia sought to identify actors who are involved in and/or could be involved in the big data for development discourse globally, with a particular focus on the 'Global South. CIS attempted to document the activities related to big data for development in and relevant to the 'Global South,' and the activities related to big data for development in the 'Global North' for the purpose of understanding actors, organizations, regions, domain trends and research gaps towards informing a big data for development research agenda. They outlined the mapping process, which included searching, screening, mapping and coding, appraisal, filling the gaps, quality assurance and synthesis. The results were coded by type of actor, type of organization, domain and region.

The initial results indicated that university work on big data largely revolved around research whilst governments generally played the role of policy actors. Civil society actors involved in the big data space were a mix of both policy actors and researchers. The exercise also revealed that a large share of data providers appeared to originate from the 'Global North' (though sometimes

operate in the 'Global South') raising important question about data ownership and ethics of data collection and use. There were also big data for development (BD4D) collaborations in partnership between 'Global North' and 'Global South' actors with businesses incorporated in developed economies are supporting and initiating BD4D projects and research in developing and/or emerging economies. CIS outlined that the two key challenges they faced when conducting the mapping exercise were language barriers (particularly for the Latin American regions), and coding individuals/organizations that fit in multiple domains and actors.

Some of the observations based on the preliminary results include (1) big data and rights dialogue can be brought into the big data and development work; (2) there is a need for grassroots initiatives around big data for development in the Global South that can draw insights from Global North institutions but are not driven by Global North institutions and (3) many actors in big data for development in the Global South are emerging. This presents an important opportunity for 'on the ground' networks to be formed.

Session 3: Discussion of way forward - Modalities of developing a research and policy agenda for the Global South

The purpose of the session was for the participants to brainstorm and identify research ideas that were at the frontiers of research in big data for development and provide suggested approaches for the development of a research agenda.

The importance of capacity building in the proposed network was underscored, particularly if the objective was to develop voices in the Global South based on the argument that people need to be given the ability to undertake the data science and there should be deep contextual understanding of marginalization, privacy and competition issues among others. Participants suggested that a network that facilitated a team-based approach to research was more beneficial than one focusing on individual researchers conducting their own research.

Within this context, participants mentioned that the focus of a network and its goals were dependent on the funder(s)'s time horizon and investment. This is particularly important given that capacity building in data science was a medium to long-term process. A pipeline would be needed that would take students in their undergraduate years to work with researchers for a couple of years and then they either go to industry or ideally policy or continue their education. Thus if the time horizon was around three years, instead of building a research pipeline, it would make more sense to focus on policy. Moreover, there could be a research focus and a policy focus. The research angle could look at research about big data for development as well as research on doing big data for development.

Another suggestion was the network to act as more of a hub where not all members of the network would collaborate on research at a given point, but rather it would serve as a platform to connect research organizations who could then choose to collaborate on research. Participants suggested to

build institutional capacity, which would spill onto national capacity to influence legislation and policy and raise the question of sustaining the network.

Similarly, looking at similar networks to gauge what has worked and what hasn't would also be beneficial. Capacity building should lead into a relationship factor thus it would be important to look at capacity in terms of institutional presence, relationship and technical. With regards to connections, if there is research being conducted on the same issue in parallel, this initiative can ensure that there would be a process of osmosis between those two projects.

APPENDIX

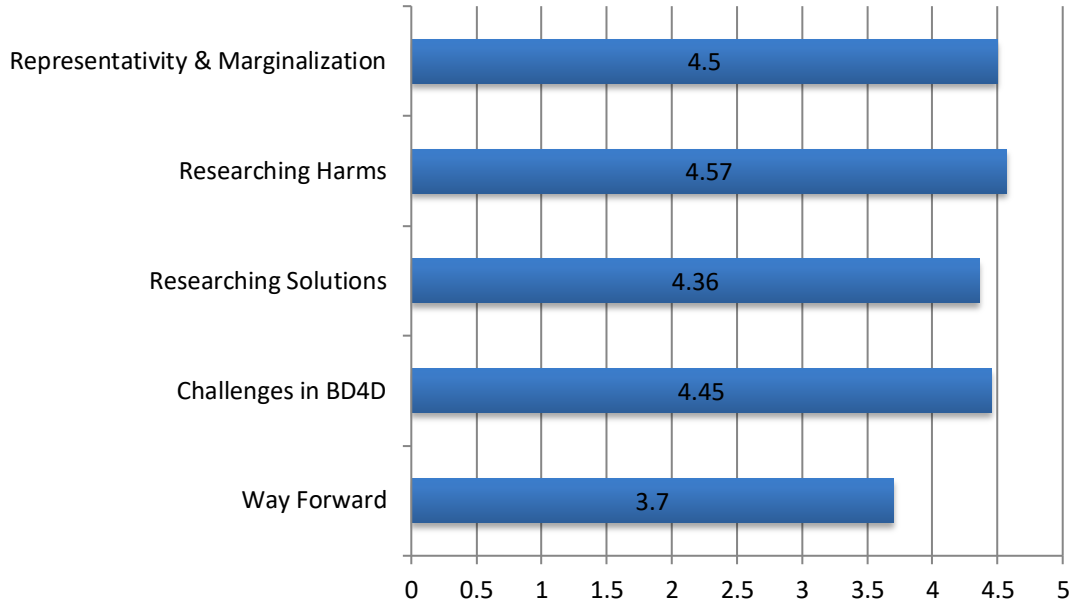
List of Participants

- Prof. Joshua Blumenstock, PhD, Assistant Professor, School of Information, University of California, Berkeley, USA
- Mr. Sumandro Chattapadhyay, Research Director – Centre for Internet and Society, India
- Ms. Katie Clancy, Program Management Officer, IDRC, Canada
- Dr. Tom Fisher, PhD, Research Officer, Privacy International
- Ms. Thavisha Gomez, Research Manager, LIRNE*asia*, Sri Lanka
- Ms. Elonnai Hickok, Director, Internet Governance – Centre for Internet and Society, India
- Mr. Sriganesh Lokanathan, Team Leader – Big Data Research, LIRNE*asia,* Sri Lanka
- Dr. Miguel Luengo-Oroz, PhD, Chief Data Scientist, UN Global Pulse, New York
- Mr. Sean McDonald, Chief Executive Officer, FrontlineSMS, USA
- Dr. Maurice McNaughton, PhD, Director, Centre of Excellence, University of the West Indies, West Indies
- Prof. Bitange Ndemo, PhD, xx, University of Nairobi, Kenya
- Dr. Juan Pane, PhD, Researcher, Latin America Open Data Initiative, Paraguay
- Dr. Fernando Perini, PhD, Senior Program Officer, IDRC, Canada
- Ms. Diastika Rahwidiati, Chief Technical Adviser, Pulse Lab Jakarta, Indonesia
- Mr. Ali Rebaie, Data Science Anthropologist, Data Aurora, Lebanon
- Dr. John Quinn, PhD, Data Scientist, Pulse Lab Kampala, Uganda
- Mr. Daniel Rodriguez, Outreach and Partnerships Coordinator, CEPEI, Colombia
- Prof. Rohan Samarajiva, PhD, Founding Chair, LIRNE*asia*, Sri Lanka
- Dr. Ruhiya Seward, PhD, Senior Programme Officer, IDRC, Canada
- Dr. Fabrizio Scrollini, PhD, Lead Researcher, Latin American Initiative for Open Data, Uruguay
- Prof. Ryosuke Shibasaki, PhD, Professor, Centre for Spatial Information Science, University of Tokyo, Japan
- Ms. Katherine Townsend, USAID
- Dr. Mario Viola, PhD, ITS Rio, Brazil

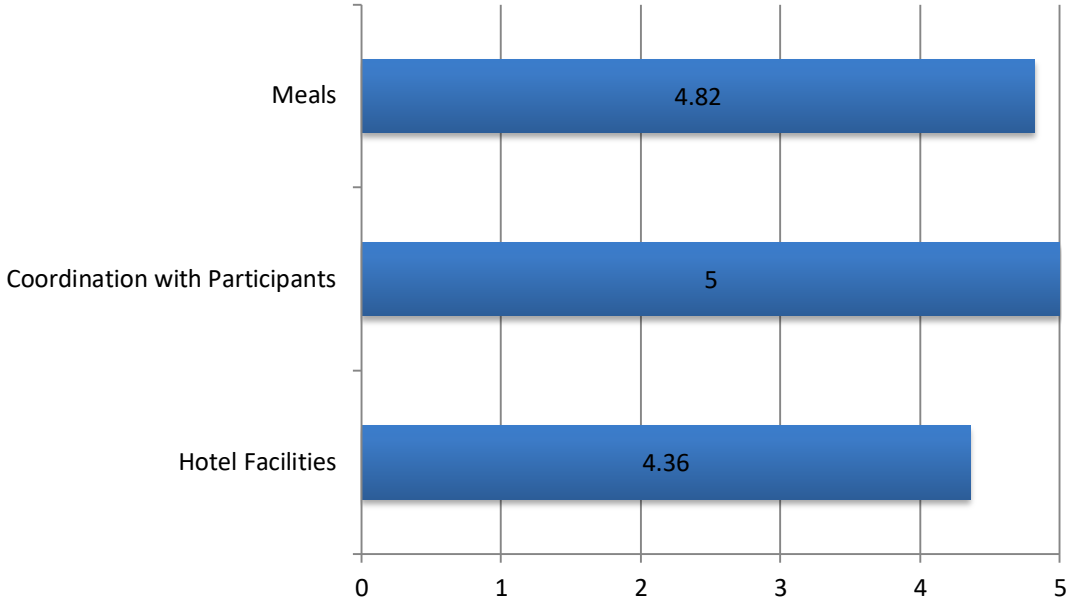
Evaluations

The workshop evaluations can be seen below. The rating was based on a five-point scale ranging from 1-abysmal to 5-excellent. All the topics, speakers and logistical arrangements were rated as being satisfactory.

Discussion Ratings



Logistics Ratings


Meals
4.82
Coordination with Participants
5
Hotel Facilities
4.36
0
1
2
3
4
5